



Tersedia Online : <http://e-journals.unmul.ac.id/>

ADOPSI TEKNOLOGI DAN SISTEM INFORMASI (ATASI)

Alamat Jurnal : <http://e-journals2.unmul.ac.id/index.php/atasi/index>



Penerapan Model BERT pada *Chatbot* dalam Platform E-Commerce

Muhammad Abdillah Rahmat ^{1)*}, Karmila ²⁾, Husain Muhammad Khatami ³⁾,
Muhammad Alief Fahdal Imran Oemar ⁴⁾

Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin

E-Mail : abdi.rahmat@unhas.ac.id¹⁾; karmila24d@ms.unhas.ac.id²⁾; khatamihm24d@ms.unhas.ac.id³⁾;
alieffahdal@unhas.ac.id⁴⁾;

ARTICLE INFO

Article history:

Received : 22 May 2025

Revised : 23 May 2025

Accepted : 29 May 2025

Available online : 10 June 2025

Keywords:

Chatshop

BERT

BLEU

ROUGE

Kata Kunci :

Chatshop

BERT

BLEU

ROUGE

APA style in citing this article:

M. A. Rahmat, Karmila, H. M.

Khatami, and M. A. F. I.

Oemar, "Penerapan Model

BERT pada *Chatbot* dalam

Platform E-Commerce," *ATASI*

: Adopsi Teknologi dan Sistem

Informasi, vol. 4, no. 1, pp. 72-

79, 2025.

[https://doi.org/10.30872/atasi.v](https://doi.org/10.30872/atasi.v4i1.3039)

[4i1.3039](https://doi.org/10.30872/atasi.v4i1.3039)

ABSTRACT

This study develops *Chatshop*, a BERT-based e-commerce chatbot designed to assist customers in searching for pizza menus. The BERT model was chosen because of its ability to understand sentence context bidirectionally, thereby increasing the accuracy in detecting user intent. This *Chatshop* allows users to find menus, get recommendations, and access price and stock availability information in real time. The evaluation was carried out using BLEU and ROUGE-L metrics, with a ROUGE-L F1 score of 32.91%. These results indicate that the chatbot is able to handle simple interactions well, but still needs improvement in answering more complex questions accurately and completely.

ABSTRAK

Penelitian ini mengembangkan *Chatshop*, sebuah *chatbot e-commerce* berbasis BERT yang dirancang untuk membantu pelanggan dalam pencarian menu pizza. Model BERT dipilih karena kemampuannya memahami konteks kalimat secara *bidirectional*, sehingga meningkatkan akurasi dalam mendeteksi *intent* pengguna. *Chatshop* ini memungkinkan pengguna menemukan menu, memperoleh rekomendasi, serta mengakses informasi harga dan ketersediaan stok secara *real-time*. Evaluasi dilakukan menggunakan metrik BLEU dan ROUGE-L, dengan hasil skor ROUGE-L F1 sebesar 32.91%. Hasil ini menunjukkan bahwa *chatbot* mampu menangani interaksi sederhana dengan baik, namun masih perlu penyempurnaan dalam menjawab pertanyaan yang lebih kompleks secara akurat dan lengkap.

2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

1. PENDAHULUAN

Industri perdagangan elektronik (*e-commerce*) berkembang pesat seiring dengan berkembangnya teknologi, termasuk algoritma rekomendasi yang dipersonalisasi dan teknologi web 2.0. Pertumbuhan industri ini menggarisbawahi pentingnya berinovasi untuk tetap kompetitif dan meningkatkan pengalaman pelanggan serta efisiensi operasional (Singh & Sai Vijay, 2024). Dengan kemajuan teknologi kecerdasan buatan (AI), robot layanan profesional memasuki pasar global dengan kecepatan yang belum pernah terjadi sebelumnya, dengan *chatbot* muncul sebagai aplikasi yang sangat penting (Lu & Zhang, 2025).

Chatbot percakapan menjadi semakin penting dalam meningkatkan pengalaman pelanggan dan meningkatkan pendapatan pada platform *E-commerce*. Alat-alat canggih ini menawarkan berbagai macam layanan, termasuk dukungan pelanggan, rekomendasi produk yang dipersonalisasi, pemberitahuan diskon promosi, dan strategi untuk pelanggan retensi (Wong et al., 2024). Dalam *e-commerce*, *chatbot* dapat membantu menjawab pertanyaan pelanggan dengan segera tanpa melibatkan manusia mana pun. Kemampuan ini dapat membantu meningkatkan kualitas layanan pelanggan dan alokasi sumber daya manusia (Ngai et al., 2021).

Dalam penelitian oleh Sanjaya, dilakukan tinjauan sistematis terhadap implementasi *chatbot* dalam bidang komersial. Studi ini menganalisis berbagai pendekatan, teknologi, dan tantangan yang dihadapi dalam penerapan *chatbot* untuk mendukung interaksi pelanggan, transaksi, dan layanan dalam sektor *e-commerce*. Hasil kajian

*) Corresponding Author

<https://doi.org/10.30872/atasi.v4i1.3039>

2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

memberikan gambaran menyeluruh terkait efektivitas *chatbot* dalam meningkatkan pengalaman pengguna dan efisiensi bisnis (Sanjaya et al., 2023).

Dalam penelitian oleh Roumeliotis, dilakukan analisis komparatif antara model GPT dan LLaMA untuk evaluasi ulasan produk dalam *e-commerce*. Model ini dievaluasi berdasarkan kemampuan memahami sentimen dan aspek ulasan. Hasilnya kedua model menunjukkan keunggulan masing-masing dalam akurasi dan efisiensi pemrosesan, yang bermanfaat untuk meningkatkan kualitas layanan pelanggan (Roumeliotis, Tselikas, & Nasiopoulos, 2024).

Di sisi lain, Habbat melakukan analisis sentimen pada data tidak seimbang dengan memanfaatkan model BERT dan teknik *ensemble stacking*. Metode ini terbukti mampu meningkatkan akurasi dan mengatasi ketidakseimbangan data secara efektif (Habbat et al., 2023).

Chatbot percakapan kini menjadi komponen krusial dalam industri *e-commerce* modern, karena kebutuhan untuk menyediakan layanan pelanggan yang aktif 24/7 dengan respons cepat dan personal tanpa campur tangan manusia. Pendekatan berbasis *rule-based* sering kali gagal menghadapi variasi struktur kalimat dan konteks pengguna, sehingga menuntut model yang lebih canggih. Keunggulan BERT dalam menangkap konteks bidirectional dan menghasilkan representasi kalimat yang kaya dan sebagaimana dibuktikan efektivitasnya dalam berbagai tugas NLP dan menjadikannya solusi ideal untuk meningkatkan akurasi deteksi *intent*. Dengan memanfaatkan kekuatan tersebut, penelitian ini merancang *Chatshop e-commerce* berbasis BERT khusus untuk pencarian menu pizza, sehingga pelanggan dapat dengan mudah menemukan menu, menerima rekomendasi sesuai preferensi, serta memperoleh informasi stok dan harga secara *real-time*.

2. TINJAUAN PUSAKA

Perkembangan pesat industri *e-commerce* dalam dekade terakhir tidak dapat dilepaskan dari peran teknologi personalisasi dan adopsi Web 2.0, yang menghadirkan interaktivitas dan pengalaman pengguna yang lebih dinamis. Menurut Singh dan Sai Vijay (2024), keberhasilan platform *e-commerce* sangat ditentukan oleh strategi berbasis teknologi digital, khususnya dalam hal personalisasi layanan, efisiensi logistik, serta peningkatan interaksi pelanggan. Di antara teknologi tersebut, *chatbot* berbasis AI menonjol sebagai solusi layanan pelanggan yang efisien, responsif, dan ekonomis.

Chatbot modern tidak lagi hanya menjawab pertanyaan statis, tetapi telah berevolusi menjadi sistem percakapan cerdas yang mampu memahami konteks, menyusun respons alami, dan bahkan memprediksi kebutuhan pengguna berdasarkan interaksi sebelumnya (Lu & Zhang, 2025). Teknologi ini memungkinkan platform *e-commerce* memberikan dukungan 24/7, mengurangi beban kerja manusia, serta meningkatkan loyalitas pelanggan. Seperti dijelaskan oleh Wong et al. (2024), *chatbot* saat ini menjalankan berbagai fungsi seperti rekomendasi produk berbasis preferensi pengguna, pengiriman notifikasi promosi, hingga keterlibatan dalam proses retensi pelanggan.

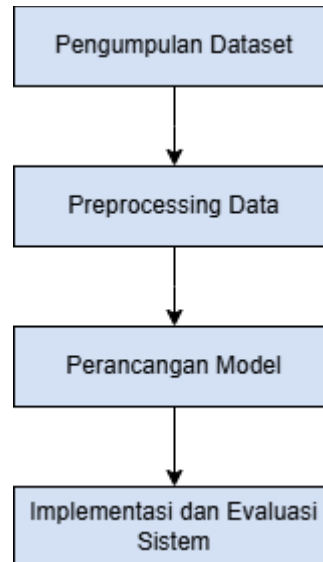
Secara arsitektural, *chatbot* dapat dikategorikan ke dalam dua model utama: *rule-based* dan *machine learning-based*. Meskipun *rule-based* cukup efektif untuk skenario sederhana, keterbatasannya dalam menangani bahasa alami dan konteks yang kompleks membuat model ini tidak lagi memadai dalam skenario komersial modern. Oleh karena itu, penelitian kini beralih ke model pembelajaran mendalam seperti BERT, GPT, dan LLaMA yang menawarkan pemrosesan konteks dua arah (*bidirectional*) dan pemahaman semantik yang mendalam (Roumeliotis et al., 2024; Devlin et al., 2018).

Dalam studi teknis, Sanjaya et al. (2023) memberikan klasifikasi sistematis terhadap implementasi *chatbot* dalam platform *e-commerce*, termasuk metrik performa seperti akurasi deteksi *intent*, waktu respons, serta kepuasan pengguna. Mereka juga menyoroti tantangan seperti integrasi lintas sistem, perlindungan privasi data, dan adaptasi terhadap variasi linguistik pengguna. Penggunaan BERT (Bidirectional Encoder Representations from Transformers) sebagai fondasi *chatbot* terbukti memberikan hasil yang menjanjikan. Habbat et al. (2023) menunjukkan bahwa dengan menggabungkan BERT dan teknik *ensemble stacking*, sistem klasifikasi sentimen mampu mengatasi masalah data tidak seimbang, yang umum ditemukan dalam data ulasan pelanggan. Dalam konteks *chatbot*, keunggulan BERT dalam memahami konteks kalimat dan makna kata berdasarkan posisi relatifnya menjadikannya solusi tepat untuk mendeteksi maksud (*intent detection*) dan mengekstraksi informasi secara otomatis.

Lebih lanjut, karya Zhang et al. (2023) menunjukkan bahwa integrasi BERT ke dalam sistem *chatbot e-commerce* memungkinkan interaksi seperti pencarian produk, pengecekan stok, dan pemberian rekomendasi berdasarkan preferensi yang diekstrak dari histori percakapan. Hal ini meningkatkan kepuasan pelanggan secara signifikan dibandingkan dengan pendekatan konvensional. Di sisi lain, untuk menjembatani antara performansi model dan efisiensi komputasi, penelitian oleh Lin et al. (2022) menyoroti strategi kompresi model (model pruning dan distillation) yang memungkinkan penerapan model *chatbot* berbasis transformer pada perangkat *edge* atau *mobile* dengan keterbatasan sumber daya.

Selain itu, tantangan dalam pengelolaan percakapan *multi-turn* (percakapan yang berlangsung lebih dari satu kali tanya-jawab) dibahas oleh Qiu et al. (2022), yang mengembangkan sistem *chatbot* yang mempertimbangkan konteks historis pengguna untuk meningkatkan relevansi respons. Secara keseluruhan, perkembangan teknologi *chatbot*, terutama berbasis BERT dan GPT, menunjukkan potensi besar dalam mendukung interaksi pengguna yang lebih cerdas, adaptif, dan personal di platform *e-commerce* modern.

3. METODE PENELITIAN



Gambar 1. Alur Penelitian

A. Pengumpulan Dataset

Pada tahap pengumpulan *dataset* untuk Chatshop dalam pemesanan pizza, data diklasifikasikan berdasarkan *intent* atau maksud dari pengguna saat berinteraksi. Dataset ini dipresentasikan dalam bentuk *dictionary intent_examples*, yang berisi berbagai contoh kalimat yang bisa digunakan oleh pengguna. Setiap *intent* seperti sapaan, menu, rekomendasi, harga, stok, dan sebagainya, memiliki sejumlah kalimat yang mencerminkan variasi cara pengguna menyampaikan maksud tersebut. Misalnya, *intent* menu yang memiliki contoh kalimat seperti “apa saja menu pizza?”, “daftar pizza”, dan “menunya apa aja?”, yang bertujuan untuk menanyakan daftar menu yang tersedia. Selain itu, *dictionary menu_pizza* yang bukan bagian dari dataset pelatihan *intent*, namun berfungsi sebagai sumber data pendukung yang berisi informasi tentang menu pizza, seperti nama, harga, stok, dan deskripsi. Informasi ini akan digunakan Chatshop untuk menjawab pertanyaan seputar harga, stok rekomendasi dan lainnya.

B. Pre-processing Data

Untuk memproses input dari pengguna, sistem melakukan tokenisasi menggunakan *tokenizer* dari model *transformer* (seperti *IndoBERT*).

Tahap *tokenizing/parsing* merupakan tahap pemotongan *string* input berdasarkan setiap kata yang menyusunnya. Karakter non-huruf dihilangkan dan dianggap sebagai *delimiter*. *Tokenizing* sangat berguna ketika suatu program pengolah teks membutuhkan data dalam bentuk kata yang terstruktur dan dibagi ke dalam *array* (Rahmat, Indrabayu, & Areni, 2019).

Dalam sistem ini tokenisasi dilakukan melalui fungsi *get_embedding()* yang menerima masukkan teks, kemudian mengubahnya menjadi token, serta menambahkan *padding* dan pemangkasan (*truncation*) hingga panjang maksimum tertentu.

C. Perancangan Model

Pada tahap ini, sistem dikembangkan dengan memanfaatkan pendekatan berbasis *deep learning* menggunakan arsitektur BERT (Bidirectional Encoder Representations from Transformers), yang dikenal mampu memahami konteks kalimat secara menyeluruh melalui mekanisme pemrosesan dua arah.

BERT adalah teknik pembelajaran mesin sumber terbuka untuk menangani bahasa alami (NLP). BERT menggunakan teks disekitarnya untuk menetapkan konteks guna membantu komputer dalam menginterpretasikan bahasa yang membingungkan (Habbat et al., 2023).

Model ini digunakan untuk mendeteksi *intent* dari masukan pengguna dalam sistem Chatshop dengan akurasi semantik yang lebih tinggi dibanding pendekatan berbasis *keyword* atau *rule-based*.

Proses dimulai dengan pembuatan fungsi *encode(text)*, yang bertugas mengubah masukan berupa teks menjadi representasi vektor numerik berdimensi tinggi. Fungsi ini memanfaatkan *tokenizer* bawaan dari model BERT untuk memecah kalimat menjadi token-token sub-kata, melakukan *padding* dan *truncation* agar panjang input seragam, serta mengubah token tersebut menjadi format tensor agar dapat diproses oleh model. Setelah input masuk ke dalam model BERT, dikeluarkan vektor representasi untuk setiap token dalam kalimat. Namun, hanya vektor dari token khusus yang diambil karena token ini secara khusus dirancang untuk merepresentasikan keseluruhan makna atau konteks kalimat secara utuh dalam berbagai tugas klasifikasi.

Hasil representasi vektor dari kalimat pengguna kemudian digunakan dalam fungsi *detect intent(user input)*, yang berperan dalam menentukan maksud atau *intent* dari input tersebut. Di dalam fungsi ini, vektor hasil *encode* dibandingkan dengan vektor dari berbagai contoh kalimat (*intent examples*) yang telah dipetakan sebelumnya untuk setiap jenis *intent*. Proses pembandingannya dilakukan dengan menghitung *cosine similarity*, yaitu sebuah metode matematis untuk mengukur tingkat kemiripan sudut antar dua vektor dalam ruang berdimensi tinggi. Semakin tinggi

skor *similarity*, semakin dekat makna dari dua kalimat tersebut. Untuk setiap iterasi, sistem menyimpan nilai kemiripan tertinggi dan *intent* yang sesuai sebagai kandidat terbaik. Jika skor *similarity* tidak melampaui ambang batas tertentu, maka sistem akan mengembalikan label “tidak_diketahui” sebagai indikasi bahwa input pengguna tidak sesuai dengan *intent* yang dikenali.

Melalui rancangan ini, model mampu mengenali maksud pengguna secara fleksibel dan kontekstual, bahkan jika struktur kalimat, pilihan kata, atau urutan kata mengalami variasi. Pendekatan ini memberikan keunggulan dalam hal generalisasi dan pemahaman bahasa alami, yang sangat penting dalam membangun sistem chat generatif yang responsif dan cerdas.

4. HASIL DAN PEMBAHASAN

A. Implementasi Sistem

Chat generatif ini dirancang untuk melayani kebutuhan pelanggan dalam mencari menu pizza secara otomatis melalui pendekatan *intent detection* berbasis pemrosesan bahasa alami. Sistem ini bertujuan untuk meningkatkan efisiensi layanan pelanggan dalam konteks *e-commerce* makanan cepat saji dengan memberikan pengalaman percakapan yang responsif, informatif, serta personal.

Struktur utama dari sistem percakapan ini dibangun di atas fungsi *Chatshop_response(user input)* yang bertanggung jawab dalam menangani seluruh logika interaksi sistem. Fungsi ini menerima input berupa teks dari pengguna, kemudian secara langsung mengintegrasikan fungsi *detect_intent(user input)* yang telah dirancang untuk mengidentifikasi maksud atau tujuan utama dari kalimat pengguna. *Intent* yang dideteksi meliputi berbagai jenis interaksi dasar seperti “sapaan”, “menu”, “rekomendasi”, “harga”, “stok”, “promo”, “ucapan terima kasih”, “identitas Chatshop”, hingga “pesan” dan “tidak diketahui”. Setiap *intent* tersebut dipetakan ke dalam blok percabangan *if-elif* yang menyimpan berbagai skenario respons yang sesuai dengan maksud pengguna.

Sebagai contoh, jika *intent* yang dikenali adalah sapaan, maka sistem tidak hanya memberikan balasan secara generik, namun juga memperhatikan konteks waktu berdasarkan kata kunci seperti “pagi”, “siang”, atau “malam”.

Untuk *intent* menu, Chatshop akan menampilkan daftar seluruh pizza yang tersedia. Proses ini dilakukan dengan mengambil seluruh kunci (*keys*) dari struktur data *menu_pizza*, lalu menggabungkannya menjadi sebuah *string* yang informatif. Sementara itu, untuk *intent* harga, sistem dilengkapi dengan kemampuan untuk mendeteksi apakah pengguna menyebutkan nama pizza tertentu. Jika ya, maka sistem akan merespons dengan menyajikan harga spesifik dari pizza tersebut. Namun, jika nama pizza tidak disebutkan secara eksplisit, maka Chatshop secara otomatis akan menampilkan daftar harga lengkap dari seluruh menu, termasuk format harga yang dirancang agar mudah dibaca oleh pengguna.

Demikian pula pada *intent* stok, sistem akan mencocokkan nama pizza yang disebutkan oleh pengguna dengan data pada *menu_pizza*. Jika ditemukan, sistem akan menampilkan jumlah stok aktual. Apabila pengguna tidak menyebutkan pizza tertentu, maka sistem akan mengembalikan informasi stok dari seluruh item menu. Hal ini mencerminkan fleksibilitas Chatshop dalam merespons variasi gaya bertanya dari pengguna, baik eksplisit maupun implisit.

Selain fitur-fitur utama tersebut, sistem juga mengakomodasi *intent* seperti rekomendasi, dimana sistem ini secara otomatis menyarankan menu populer seperti “Pizza Keju”. *Intent* lainnya seperti promo memberikan informasi tentang status promosi saat ini; identitas Chatshop memberikan informasi mengenai peran dan fungsi Chatshop; sedangkan ucapan terima kasih dan pesan menangani percakapan sopan dan perintah pemesanan secara langsung. Untuk masukan yang tidak dapat dipetakan ke dalam *intent* manapun, sistem akan merespons dengan pesan kesalahan yang sopan dan informatif, yaitu “Maaf, saya tidak mengerti. Bisa ulangi dengan kata yang berbeda?”, yang menunjukkan bahwa Chatshop ini dirancang untuk mengantisipasi ketidakpastian input pengguna secara manusiawi.

Salah satu keunggulan dari implementasi ini adalah keberadaan fitur pencatatan atau *logging* percakapan yang terintegrasi secara langsung melalui struktur data *chat log*. Setiap percakapan yang terjadi akan disimpan dalam bentuk *dictionary* yang memuat tiga informasi penting, yaitu input dari pengguna, *intent* yang berhasil dideteksi, serta respons yang diberikan oleh Chatshop. Setelah sesi percakapan berakhir (yakni saat pengguna mengetik kata kunci seperti “keluar”, “exit”, atau “quit”), sistem akan mengeksekusi fungsi *save_log_to_csv()* untuk menyimpan seluruh interaksi tersebut ke dalam file *chat_log.csv*. File ini dapat digunakan untuk berbagai kebutuhan sekunder seperti analisis performa Chatshop, pelatihan ulang model, atau evaluasi kualitas deteksi *intent* di masa mendatang.

Secara arsitektural, sistem ini dibangun secara modular dengan pemisahan yang jelas antara fungsi deteksi *intent*, logika respons, dan pencatatan log interaksi. Pola perulangan *while true* digunakan sebagai mekanisme utama untuk menjalankan sesi percakapan interaktif dalam mode terminal, di mana pengguna dapat memasukkan input secara bebas dan mendapatkan respons langsung dari sistem secara *real-time*.

```
👋 Selamat datang di ChatShop Pizza 🍕
Ketik 'keluar' untuk berhenti

Kamu: hai
🤖 ChatShop: Hai juga! Ada yang bisa saya bantu?
Kamu: ada menu apa saja?
🤖 ChatShop: Kami punya menu Pepperoni, Keju, Daging Sapi, Ayam BBQ, Vegetarian.
Kamu: daftar harga?
🤖 ChatShop: Daftar harga pizza adalah:
- Pepperoni: Rp50000
- Keju: Rp40000
- Daging Sapi: Rp55000
- Ayam BBQ: Rp52000
- Vegetarian: Rp45000

Kamu: paling laris apa?
🤖 ChatShop: Pizza Keju adalah yang paling laris! Mau coba?
Kamu: oke, terima kasih
🤖 ChatShop: Terima kasih kembali
Kamu: keluar
Bot: Sampai jumpa! Terima kasih 😊
```

Gambar 2. Tampilan Percakapan

Gambar diatas menampilkan simulasi percakapan antara pengguna dan sebuah chat pemesanan pizza yang berbahasa Indonesia. Chatshop ini dirancang untuk memberikan pengalaman interaktif dalam memesan pizza dengan fitur-fitur seperti memberikan informasi menu, daftar harga, rekomendasi pizza, serta stok yang tersedia. Sistem ini merespons setiap input pengguna dengan ramah dan informatif. Ini menunjukkan bahwa chat generatif ini tidak hanya memberikan jawaban fungsional, tetapi juga dirancang agar terasa lebih manusiawi dan menyenangkan bagi pengguna.

Dari segi fungsionalitas, sistem ini cukup lengkap untuk kebutuhan dasar pemesanan makanan. Ia mampu menawarkan pilihan menu, memberikan saran, menzampilkan harga, dan memeriksa stok secara *real-time*. Percakapan ditutup dengan perintah "keluar", yang juga dikenali oleh sistem sebagai sinyal untuk mengakhiri interaksi. Penggunaan bahasa yang sopan dan akrab membuat *chatbot* ini cocok untuk pengguna umum. Secara keseluruhan, Chatshop ini menunjukkan desain percakapan yang baik, *user-friendly*, dan mendukung proses pemesanan dengan efektif.

B. Evaluasi Sistem

Dalam evaluasi kualitas respons Chatshop, ada dua metrik utama dari bidang NLP, yaitu BLEU dan ROUGE-L. BLEU (Bilingual Evaluation Understudy) adalah salah satu metrik evaluasi pertama yang digunakan secara luas untuk menilai kualitas teks yang dihasilkan oleh sistem terjemahan mesin atau model bahasa lainnya. BLEU membandingkan n-gram dari hasil keluaran model dengan satu atau beberapa referensi (jawaban ideal) dan menghitung kesamaan berdasarkan tumpang tindih kata (Wentzel, 1922). Sedangkan ROUGE (Recall-Oriented Understudy for Gisting Evaluation) adalah kumpulan metrik yang digunakan untuk mengevaluasi sistem otomatis dalam menghasilkan ringkasan teks dan jawaban, seperti pada tugas *summarization* dan chatbot response. Salah satu variasinya, ROUGE-L, mengukur kesamaan berdasarkan *Longest Common Subsequence* (LCS) antara *output* dan referensi (Lin, 2004).

Dalam sistem ini BLEU digunakan untuk mengukur kesamaan antara kalimat keluaran Chatshop (*user input*) dengan kalimat referensi (*response*) berdasarkan n-gram yang cocok, dengan tambahan *smoothing* untuk hasil yang lebih stabil pada kalimat pendek. Sementara itu, ROUGE-L menilai seberapa banyak urutan kata terpanjang yang cocok antara respons Chatshop dan referensi, dan menghitung *precision*, *recall*, serta *F1-score* dari kesamaan tersebut.

Setiap baris dalam *file log* interaksi *chat_log.csv* dievaluasi secara individu, dan hasil skor BLEU serta ROUGE-L ditampilkan per interaksi. Selain itu, nilai BLEU untuk seluruh interaksi divisualisasikan dalam grafik garis untuk memantau performa *chat* generatif ini dari waktu ke waktu. Evaluasi ini membantu mengukur seberapa baik chat generatif ini memahami dan membalas masukan pengguna dengan relevansi dan ketepatan yang mendekati respon ideal.

Untuk mengetahui sejauh mana kualitas respon chat generatif ini dalam menanggapi masukan dari pengguna, dilakukan evaluasi menggunakan dua metrik NLP yaitu BLEU dan ROUGE-L. Metrik ini membantu mengukur kemiripan antara jawaban yang diberikan Chatshop dan jawaban referensi ideal secara kuantitatif. Evaluasi dilakukan terhadap setiap interaksi yang terekam dalam *log* percakapan. Percakapan yang digunakan adalah percakapan yang ada pada gambar 3.1 tadi.

Tabel 1. Evaluasi *Blue* dan *Rouge-l* dalam Interaksi 1

User	Nilai
User	Hai
Chatshop	Hai juga! Ada yang bisa saya bantu?
BLEU	0.0001
Rouge-L precision	1.0000
Rouge-L Recall	0.1429
Rouge-L F1-Score	0.2500

Pada interaksi ini, meskipun makna balasan dari Chatshop sangat baik dan ramah, BLEU yang sangat rendah menunjukkan bahwa susunan kata dari *output* sangat berbeda dengan referensi (mungkin hanya "hai"). *ROUGE-L precision* tinggi (1.0) karena semua kata referensi (kata "hai") ada dalam balasan. Namun, *recall* rendah (0.1429)

karena balasan jauh lebih panjang dari referensi. Nilai *F1- Score* sebesar 0.2500 merupakan rata-rata harmonik dari *precision* dan *recall* yang mencerminkan ketidakseimbangan antara keduanya, meskipun kata-kata cocok (*precision* tinggi), terlalu banyak informasi tambahan membuat *recall* rendah.

Tabel 2. Evaluasi *Bleu* dan *Rouge-l* dalam Interaksi 2

User	Nilai
User	Ada menu apa saja?
Chatshop	Kami punya menu pepperoni, keju, daging sapi, ayam BBQ, vegetarian.
BLEU	0.0073
Rouge-L precision	0.2500
Rouge-L Recall	0.1000
Rouge-L F1- Score	0.2500

Interaksi ini merupakan respons yang diberikan Chatshop mencakup semua menu, namun BLEU dan ROUGE-L menunjukkan performa rendah secara tekstual. BLEU rendah karena susunan kata dan frasa kemungkinan besar berbeda dengan referensi. *ROUGE-L precision* (0.2500) menunjukkan bahwa hanya sebagian kecil kata referensi cocok dengan *output*. *ROUGE-L recall* lebih rendah (0.1000), mengindikasikan banyak kata penting dalam referensi yang tidak dikenali di jawaban. Nilai *F1-Score* sebesar 0.1429 memperlihatkan bahwa meskipun informasi tersampaikan, kesamaan literalnya masih lemah menurut metrik otomatis.

Tabel 3. Evaluasi *Bleu* dan *Rouge-l* dalam Interaksi 3

User	Nilai
User	Daftar harga?
Chatshop	Daftar harga pizza adalah: - Pepperoni: Rp50000 - Keju: Rp40000 - Daging Sapi: Rp55000 - Ayam BBQ: Rp52000 - Vegetarian: Rp45000
BLEU	0.0001
Rouge-L precision	1.0000
Rouge-L Recall	0.1250
Rouge-L F1- Score	0.2222

Walau BLEU mendekati nol, *ROUGE-L precision* sempurna (1.0) menunjukkan bahwa kata-kata dari referensi seluruhnya muncul di jawaban. Namun, *recall* rendah di 0.1250 karena referensi mengandung informasi tambahan atau format berbeda yang tak tertangkap dalam *output*. Nilai *F1-Score* sebesar 0.2222 menandakan bahwa ada kesenjangan antara kesesuaian kata dan kelengkapan informasi. Meski jawaban sebenarnya lengkap, struktur daftar dan format menyebabkan evaluasi literal menjadi tidak adil bagi jenis respons ini.

Tabel 4. Evaluasi *Bleu* dan *Rouge-l* dalam Interaksi 4

User	Nilai
User	Paling laris apa?
Chatshop	Pizza Keju adalah yang paling laris! Mau coba?
BLEU	0.0420
Rouge-L precision	0.6667
Rouge-L Recall	0.2500
Rouge-L F1- Score	0.3636

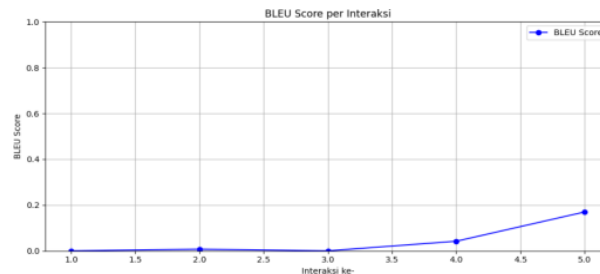
Balasan *Chatshop* sudah sesuai secara makna dan terasa natural, dengan tambahan gaya promosi (“Mau coba?”). BLEU meningkat karena frasa penting cocok. *ROUGE-L precision* menunjukkan bahwa 2/3 kata dari referensi muncul dalam *output*, namun *recall* tetap rendah karena sebagian referensi tidak tertangkap. Nilai *F1-Score* 0.3636 menunjukkan bahwa meskipun ada kecocokan kata, keluaran masih belum menutupi semua kata penting dari referensi. Hal ini mencerminkan timbal balik antara keakuratan dan kelengkapan dalam respons.

Tabel 5. Evaluasi *Bleu* dan *Rouge-l* dalam Interaksi 4

User	Nilai
User	Oke, terimakasih
Chatshop	Terimakasih kembali
BLEU	0.1699
Rouge-L precision	0.6667

User	Nilai
ORouge-L Recall	0.6667
Rouge-L F1- Score	0.6667

Dalam tabel interaksi terakhir ini menghasilkan nilai evaluasi tertinggi. BLEU relatif tinggi karena frasa “terima kasih” sama persis. *Precision* dan *recall* ROUGE-L sama- sama 0.6667, menandakan bahwa sebagian besar kata dan struktur sesuai dengan referensi. Nilai *F1-Score* sebesar 0.6667 menunjukkan bahwa respons ini seimbang dalam hal kecocokan kata dan cakupan informasi. Interaksi singkat dan formal seperti ini cenderung lebih mudah untuk dinilai secara positif oleh metrik otomatis, karena memiliki sedikit variasi linguistik dan struktur.



Gambar 3. Grafik skor BLEU dalam Interaksi

Grafik *BLEU score* per interaksi di atas menunjukkan tren perubahan performa Chatshop dalam lima sesi percakapan, diukur menggunakan metrik BLEU yang menilai kesamaan urutan kata (n-gram) antara jawaban Chatshop dan referensi yang ideal. Pada tiga interaksi pertama, *BLEU score* sangat rendah mendekati nol, mengindikasikan bahwa susunan kata dalam respons sistem sangat berbeda dari referensi, meskipun mungkin secara makna masih bisa diterima. Ini wajar terjadi ketika jawaban Chatshop lebih panjang atau memiliki variasi linguistik yang tinggi. Pada interaksi keempat, skor mulai meningkat karena jawaban sistem mulai mengandung n-gram yang cocok dengan referensi, meskipun belum terlalu signifikan. Peningkatan yang paling tajam terjadi pada interaksi kelima, di mana *BLEU score* mendekati 0.17. Hal ini mencerminkan kesamaan literal yang tinggi antara respons Chatshop dan referensi, karena percakapan pada interaksi tersebut sangat sederhana (“oke, terima kasih” dan “terima kasih kembali”), sehingga struktur kalimat yang dihasilkan mudah cocok secara tekstual. Secara keseluruhan, grafik ini menegaskan bahwa BLEU lebih akurat untuk menilai respons yang pendek dan terstruktur baku, namun kurang ideal untuk menilai jawaban fleksibel yang tetap benar secara makna.

 Akurasi Keseluruhan ChatShop (berdasarkan ROUGE-L F1): 32.91%

Gambar 4. Hasil akurasi berdasarkan Rouge-L F1

Gambar diatas menampilkan informasi bahwa akurasi keseluruhan dari Chatshop berdasarkan metrik ROUGE-L F1 adalah sebesar 32.91%. Nilai ini menggambarkan seberapa baik respons Chatshop dalam meniru struktur kalimat referensi yang ideal, dengan mempertimbangkan keseimbangan antara *precision* (jumlah kata yang cocok dibandingkan dengan total kata dalam output Chatshop) dan *recall* (jumlah kata yang cocok dibandingkan dengan total kata dalam referensi). Nilai *F1-Score* 32.91% menunjukkan bahwa secara rata-rata, Chatshop mampu menghasilkan jawaban yang memiliki sekitar sepertiga kecocokan literal terhadap referensi. Angka ini mengindikasikan performa yang masih perlu ditingkatkan, terutama dalam hal penyusunan kata dan kesesuaian kalimat. Meskipun demikian, perlu diingat bahwa ROUGE-L F1 hanya menilai kesamaan permukaan (literal), bukan pemahaman makna atau konteks, sehingga Chatshop bisa saja memberikan jawaban yang benar secara semantik namun tetap memperoleh nilai *F1* yang rendah. Maka, evaluasi tambahan berbasis *human judgement* juga tetap penting dalam menilai kualitas Chatshop secara menyeluruh.

5. KESIMPULAN DAN SARAN

A. Kesimpulan

Dalam sistem ini, hasil evaluasi menggunakan metrik BLEU dan ROUGE-L, Chatshop menunjukkan performa yang masih perlu ditingkatkan, dengan skor ROUGE-L F1 keseluruhan sebesar 32.91%. Meskipun Chatshop mampu memberikan respons yang relevan untuk interaksi sederhana seperti sapaan dan ucapan terima kasih, respons pada interaksi yang memerlukan informasi lebih kompleks seperti daftar menu dan harga masih kurang akurat dan lengkap. Hal ini terlihat dari rendahnya nilai *recall* dan *precision* pada beberapa interaksi.

Secara keseluruhan, Chatshop belum mampu secara konsisten menyajikan informasi yang sesuai dengan ekspektasi atau referensi yang diharapkan. Meskipun Chatshop ini mampu merespons interaksi sederhana dengan cukup baik, performanya menurun pada percakapan yang membutuhkan jawaban lebih kompleks dan informatif. Hal ini menunjukkan bahwa masih diperlukan perbaikan dalam hal kelengkapan informasi yang disampaikan, struktur jawaban yang lebih sistematis, serta kemampuan memahami dan menyesuaikan konteks secara lebih tepat.

Penyempurnaan di aspek-aspek tersebut diharapkan dapat meningkatkan kualitas interaksi *chatbot* secara keseluruhan dan memberikan pengalaman pengguna yang lebih optimal.

B. Saran

Berdasarkan hasil evaluasi terhadap interaksi Chatshop, terdapat beberapa hal yang dapat menjadi bahan pengembangan untuk penelitian selanjutnya. Pertama, diperlukan penyesuaian terhadap struktur dan gaya bahasa respons Chatshop agar lebih selaras dengan referensi (*ground truth*) yang digunakan dalam evaluasi berbasis metrik otomatis seperti BLEU dan ROUGE. Penyesuaian ini bertujuan untuk meningkatkan nilai evaluasi kuantitatif tanpa mengorbankan makna dan relevansi jawaban.

Kedua, disarankan untuk memperkaya variasi respons Chatshop dengan mempertimbangkan konteks dan gaya bahasa pengguna, terutama dalam situasi informal. Hal ini dapat meningkatkan pengalaman interaksi pengguna serta membuat percakapan lebih alami. Selain itu, evaluasi performa Chatshop hendaknya tidak hanya bergantung pada metrik kesamaan teks secara leksikal, tetapi juga dilengkapi dengan evaluasi berbasis kemiripan semantik atau penilaian subjektif oleh evaluator manusia. Dengan pendekatan tersebut, kualitas Chatshop dapat dinilai secara lebih holistik, mencakup aspek akurasi informasi, relevansi konteks, dan kenyamanan dalam berinteraksi.

6. DAFTAR PUSTAKA

- Singh, S., & Sai Vijay, T. (2024). Technology roadmapping for the e-commerce sector: A text-mining approach. *Journal of Retailing and Consumer Services*, 81, 103977. <https://doi.org/10.1016/j.jretconser.2024.103977>
- Lu, Y., & Zhang, J. (2025). Balancing identity diversity and product contexts: Understanding consumer trust in AI-enhanced *chatbot* services. *Journal of Retailing and Consumer Services*, 84, 104205. <https://doi.org/10.1016/j.jretconser.2024.104205>
- Wong, C. M., Feng, F., Zhang, W., Chen, H., Vong, C. M., & Chen, C. (2024). Billion-scale pre-trained knowledge graph model for conversational *chatbot*. *Neurocomputing*, 606, 128353. <https://doi.org/10.1016/j.neucom.2024.128353>
- Ngai, E. W. T., Lee, M. C. M., Luo, M., Chan, P. S. L., & Liang, T. (2021). An intelligent knowledge-based *chatbot* for customer service. *Electronic Commerce Research and Applications*, 50, 101098. <https://doi.org/10.1016/j.elerap.2021.101098>
- Sanjaya, W., Calvin, Muhammad, R., Meiliana, & Fajar, M. (2023). Systematic Literature Review on Implementation of *Chatbots* for Commerce Use. *Procedia Computer Science*, 227, 432–438. <https://doi.org/10.1016/j.procs.2023.10.543>
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. *Natural Language Processing Journal*, 6, 100056. <https://doi.org/10.1016/j.nlp.2024.100056>
- Habbat, N., Nouri, H., Anoun, H., & Hassouni, L. (2023). Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning. *Engineering Applications of Artificial Intelligence*, 126, 106999. <https://doi.org/10.1016/j.engappai.2023.106999>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Zhang, L., Sun, X., & Yu, Y. (2023). Personalized *Chatbot* System Using BERT in Online Food Ordering. *IEEE Access*, 11, 12983–12994. <https://doi.org/10.1109/ACCESS.2023.3241123>
- Lin, T., Ma, J., & Liu, J. (2022). Model Compression Techniques for Deploying *Chatbots* on Mobile Platforms. *Information Processing & Management*, 59(6), 103069.
- Qiu, M., Li, F., & Gao, Y. (2022). Multi-turn context-aware response generation in e-commerce *chatbots*. *ACM Transactions on Information Systems*, 40(4), 1–28.
- Rahmat, M. A., Indrabayu, & Areni, I. S. (2019). Hoax web detection for news in Bahasa using support vector machine. In *2019 International Conference on Information and Communications Technology (ICOIACT)* (pp. 332–336). IEEE. <https://doi.org/10.1109/ICOIACT46704.2019.8938425>
- Wentzel, G. (1922). *Funktenlinien im Röntgenspektrum*. *Annalen der Physik*, 371(23), 437–461. <https://doi.org/10.1002/andp.19223712302>
- Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)* (pp. 25–26).
- Brahi, A., Touahria, M., & Tari, A. (2019). Toward conversational recommendation systems: A comparative study. *International Journal of Computer Applications*, 177(38), 1–7. <https://doi.org/10.5120/ijca2019919641>